A Survey on Information Diffusion in Online Social Network

Abhishek Kesharwani[1],Dr. Samarendra Mohan Ghosh[2]

[1]Research Scholar, Department of Computer Science & Engineering Dr. C. V. Raman University, Kargi Road Kota, Bilaspur (C.G) Guide
[2]Dr. C. V. Raman University, Kargi Road Kota, Bilaspur (C.G)

Received: 15.07.2024 Revised: 16.08.2024 Accepted: 07.09.2024

ABSTRACT
An online social network plays a significant role in the massive dissemination of knowledge. Many efforts have been made to develop to comprehend this process, from modelling information dissemination to identifying influential spreaders to detecting hot topics. We offer a taxonomy that encapsulates the state-of-the-art in this article along with an overview of representative approaches to these problems. The goal is to offer a thorough analysis and direction for current initiatives related to knowledge dissemination via social networks. The purpose of this survey is to assist researchers in rapidly comprehending current work and potential areas for improvement.

Keywords: spreaders, dissemination, work, potential.

INTRODUCTION

Hundreds of millions of people utilize online social networks to create and access content globally. They give access to an enormous knowledge source with never-before-seen dimensions.By promoting the dissemination of fresh information and a variety of opinions, online social networks significantly contribute to the spread of knowledge [3]. They have shown to be extremely effective in a variety of circumstances, such as Twitter during the 2008 U.S. presidential elections [23] and Facebook during the 2010 Arab Spring [22]. The extraction of useful information from this massive amount of data has become a current emphasis due to the impact of online social networks on society. Social networks are dynamic environments where events, issues, interests, and so on happen quickly. It is important to capture, comprehend, visualize, and End users and scholars alike are beginning to hold predictions in high regard. This is driven by the possibility that comprehending these networks' dynamics could improve the ability to track occurrences (like evaluating revolutionary waves) and resolve problems (like prior to terrorist attacks, predicting natural disasters, enhancing corporate performance (by, for example, improving social media marketing strategies), etc. In order to capture, evaluate, extract knowledge from, and anticipate information spread in online social networks, academics have thus created a number of methodologies and models in recent years.Here, as computer scientists, we concentrate on the specific situation of information diffusion in online social networks, which poses the following queries:What information is most popular and diffuses the fastest; (ii) How, why, and through what channels is information diffusing now and will it do so in the future; (iii) Which network members are crucial to the diffusion process?Thispaper's primary objective is to provide a concise overview of the area by reviewing advances pertaining to these challenges. In light of this, we identify the advantages and disadvantages of current methods and organize them according to a taxonomy. That the purpose of this study is to provide guidelines to scientists and practitioners who plan to develop new techniques in this field. We give a library of existing approaches in this domain, which will be useful for developers who want to apply existing techniques on specific challenges.

Fundamental Information Diffusion And Online Social Networks

The utilization of a specific web-service, commonly known as an online social network (OSN), leads to social networking site (SNS), which enables users to: (i) publish messages on their profile page; (ii) connect with other users directly, thereby forming social bonds. A user-generated content system that enables communication and information sharing is, in practice, what an OSN is Formally, an OSN is represented as a graph with users as nodes and relationships as edges, which can be directed or not depending on how the SNS is configured. oversees interpersonal connections. More specifically, it depends on whether it permits bilateral (like Facebook's social model of friendship) or unilateral (like Twitter's social model of following) connections. The primary means of information in these systems are messages. Users post messages to exchange or convey different types of information, including ideas, opinions, and suggestions for products. A message can be identified by one or more of the following: (i) a text; (ii) an author; (iii) a time stamp; and, optionally, (iv) the group of individuals (referred to as "mentioned users" in social networking parlance) to whom the message is sent expressly. Figure 1 depicts anOSN enhanced by the messages shared by its four members, depicted as a

directed graph. An arc with the symbol e = (ux, uy) indicates that messages published by "uy" are visible to user "ux." According to this depiction, the user "u1" is exposed to the content that "u2" and "u3" exchange, for instance. It also suggests that nobody reads the texts that "u4" writes.

First Definition (Topic):
a logical group of phrases with similar semantic meanings that communicate a single idea. There are three practical interpretations of this definition: (i) a collection of terms S such that |S| = 1. for example, {"obama"}; (ii) a set S of terms, where |S| > 1; for example, {"obama", "visit", "china"}; and (iii) a probability distribution over a set S of terms. With one of the popular formalisms described in Definition 1, any piece of information can be turned into a subject [6, 30]. The material created by an OSN's members is a stream globally.
messages. The stream generated by the network's participants is shown in Figure 2 in the preceding example. You can think of that stream as a series of choices (e.g., whether to embrace an issue or not), with subsequent individuals observing the deeds of preceding individuals. People are therefore influenced by the things that other people do. This phenomenon is referred to as social influence [2] and is described as follows:
Defining Social Influence in Definition 2. Imitation is a social phenomena that people might experience or engage in. It refers to the idea that a user's activities can influence those in his connections to act in a similar manner. When someone "retweets" another person, for instance, influence is made clear.

Definition 3
Behaviour of the herd. a social behaviour that happens when several people act in the same way, without necessarily ignoring their signals for personal information. The Information Cascade is defined in

Definition 4
A social network activity in which users adopt information by ignoring their own signals and drawing conclusions from the behaviours of others before them.

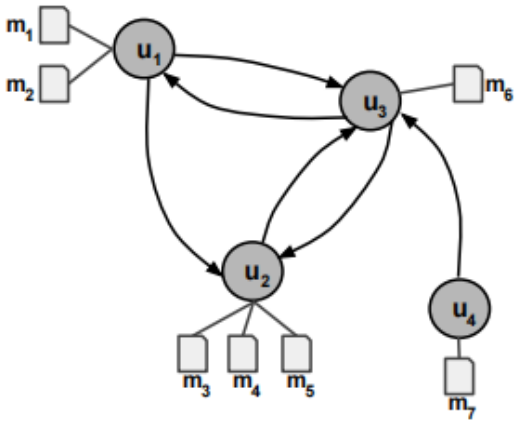

Figure 1: An example of OSN enriched by users' messages. Users are denoted u_i and messages m_j . An arc (u_x, u_y) means that u_x is exposed to the messages published by u_y.

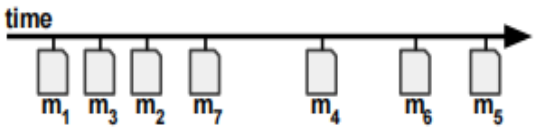

Figure 2: The stream of messages produced by the members of the network depicted on Figure 1.

Information can propagate throughout a network based on the social impact effect by using the informational cascade and herd behaviour concepts, which we define in Definitions 3 and 4, respectively.In this setting, certain subjects have the potential to gain enormous popularity, disperse around the globe, and inspire new trends. Eventually, the following can be used to summarize the components of an information diffusion process occurring in an OSN: I a bit of data delivered by messages; (ii) it disseminates over the network's edges in accordance with certain mechanics; and (iii) it depends on particular attributes of the nodes and edges. The most

pertinent new research on these many topics will be covered in the parts that follow, along with an evaluation of their advantages and disadvantages and potential areas for development.

Identifying Trending Subjects

Creating automated methods to deliver information is one of the primary goals of research on information spread. a broad perspective on the subjects that will drive the network and become popular over time. In order to summarize discussions, suggest popular topics to users, or forecast future trends in popularity, this entails extracting "tables of content."Conventional topic detection methods, which were created for the analysis of static corpora, are not suitable for the message streams produced by OSNs. It has been proposed to concentrate on bursts for the efficient detection of subjects in textual streams. Kleinberg [26] presents a state machine in his groundbreaking work to simulate the arrival timings of documents in a stream and identify bursts, presuming that all the documents belong to back to the same subject. According to Leskovec et al. [27], the ascending and descending patterns of popularity—or, to put it another way, consecutive bursts of popularity—make up the temporal dynamics of the most popular topics on social media.

An illustration of the temporal dynamics of popular subjects in OSNs is often displayed in Figure

Definition 5

(Bursty subject). a behavior related to a subject that has received a lot of attention during the time period it has been covered, but not often before or after. Here, we describe techniques intended to identify from a stream of topically diverse communications the topics that have generated bursts of attention, i.e. bursty topics (see Definition 5). Every method that follows relies on computing specific frequencies and operates with discrete data. As a result, they need to separate the message stream. Transform is used to do this.

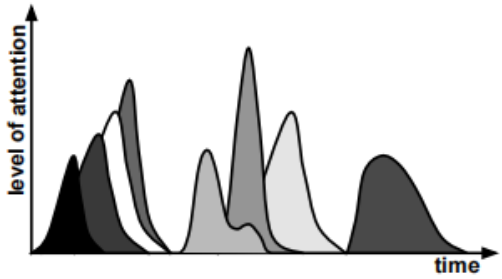

Figure 3: Temporal dynamics of popular topics. Each shade of gray represents a topicconverting the unprocessed continuous data into a series of messages that are published over uniformly sized time intervals.

Figure 4 provides an illustration of this approach by showing a potential discretization of the stream.Previously illustrated in Figure 2. Given that it establishes the level of detail in the topic detection, this pre-processing phase is not insignificant. Topics that gained popularity during brief periods of time can be identified with a very fine discretization (i.e., small time-slices), but not with a discretization that uses longer time-slices.

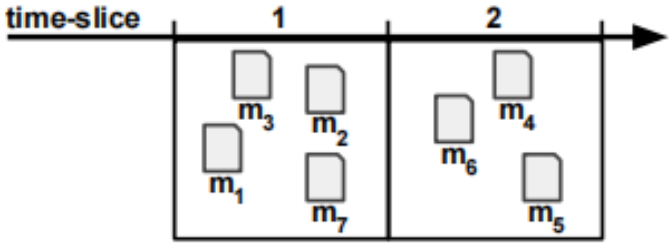

Figure 4: A possible discretization of the stream of messages shown on Figure 2.

Because it is based on a normalized term frequency measure, Shamma et al.'s [46] straightforward model, PT (i.e., Peaky Topics), is comparable to the traditional tf-idf model [44]. To determine the amount of sufficient to distinguish a topic with clarity. Consequently, more advanced techniques have been created.A non-Markov on-line LDA Gibbs sampler topic model, or OLDA, is proposed by AlSumait et al. [1] as an online topic model. Latent Dirichlet Allocation, or LDA, is essentially a statistical generative model that uses a hierarchical Bayesian network to associate words and mes-overall term usage, they view every time slice as a pseudo-

document made up of every message within the relevant collection. The definition of the normalized term frequency, or ntft, is as follows:
Ntft,i = tft,icft, where tft,i is the frequency of term t at the ith time slice and cft is the frequency of term t across the message stream. Bursty topics defined as single phrases are ranked based on that measure. But certain terminologies can be unclear or polysemous, and one phrase doesn't always seem to be guides through hidden subjects. The generating method consists of representing documents as random mixes over latent themes, where each topic is defined by a distribution over words. The goal of OLDA is to direct the learning of the new generative process by gradually updating the topic model at each time slice with the associated collection of messages and the previously generated model as a prior. This approach creates an evolutionary matrix for every topic, capturing the topic's change across time and enabling the detection of bursty topics. The Temporal and Social Terms Evaluation (TSTE) technique, which takes into account the temporal and social aspects of the message stream, is proposed by Cataldi et al. [6]. With this in mind, they create a five-step procedure that starts with formalizing the content of the messages as vectors of terms, each of which has its relative frequency calculated using the augmented normalized term frequency [43]. Next, the Page Rank algorithm and the relationships between the active authors are used to evaluate each author's authority [35]. It enables the modeling of each term's life cycle using a biological metaphor that is predicated on the computation of energy and nutrition values that take advantage of the user's authority. employing supervised or unsupervised methods based on the determination of a critical drop value the majority of bursty phrases, the suggested approach can identify the energy. Lastly, a co-occurence based metric is given to define bursty subjects as collections of phrases.
These techniques pinpoint specific subjects that have previously generated waves of interest. Lu et al.'s technique [40] allows one to forecast which In the foreseeable future, certain themes will gain attention. The authors suggest to modify the Moving Average Convergence Divergence, or MACD, technical analysis indicator, which is primarily utilized for stock price research, in order to discover single-term bursty subjects. The idea behind MACD is to create a momentum oscillator from of two trend-following indicators, specifically a shorter- and longer-term moving average of terms frequency. The difference between the long and shorter moving averages is used to determine the trend momentum. The authors provide two easy guidelines for determining when a term's trends may increase: (i) at the moment when the topic is starting to gain momentum when the trend momentum value shifts from negative to positive; conversely, when the trend momentum value shifts from positive to negative, the topic is receiving less attention. To find intriguing subjects in OSNs, the aforementioned techniques rely on identifying uncommon phrase frequencies in exchanged messages. Still, additional and more often, non-textual content including URLs, images, and videos are posted by OSNs users. Takahashi et al. [47] suggest that rather than concentrating on the textual content, non-textual content should be addressed by exploiting mentions found inside messages to identify bursty themes content. It's customary in social situations to make direct reference to messages in order to spark conversation. To do this, they create a method based on SDNML (Sequential Discounting Normalized Maximum Likelihood) change point detection along with an anomaly score. Each user's typical mentioning behavior, which is evaluated using a probability model, is taken into account while computing the anomaly.The survey methodologies are compiled on four axes in Table 1. Based on four primary criteria, the table is organized to enable a brief contrast Topic definition, dimensions included in each approach, content types handled by each method, and real or predicted burst detection/prediction are the four main questions that need to be answered. Note that the purpose of the table is to provide a worldwide comparison rather than to indicate a preference for one approach over another.

Table 1: Summary of topic detection approaches w.r.t topic definition, incorporated dimensions, handled content and the task.

reference	topic definition			dimension(s)		content type		task type	
	single term	set of terms	distribution	content	social	textual	non-textual	observation	prediction
PT	x			x		x		x	
OLDA			x	x		x		x	
TSTE		x		x	x	x		x	
SDNML	x				x	x	x	x	
MACD	x			x		x			x

4. Information Diffusion Modeling

For evaluating the propagation of misinformation, halting the transmission of viruses, and other applications, modeling the dissemination of information is quite interesting. We first go over the fundamentals of diffusion modeling in this section and after that describe the various models put forth to depict or forecast the spreading processes in OSNs."Activation Sequence," defined in

Definition 6

arranged nodes that represent the sequence in which a piece of information was adopted by the network's nodes.

Definition 7

(Cascade Spreading). a directed tree with the initial activation sequence node as its root. The tree reveals in the same order as the activation sequence, capturing the influence between nodes (branches signify to whom the information was distributed).

The diffusion process can be divided into two categories: its temporal dynamics, or how the diffusion rate changes over time, and its structure, or the diffusion graph that shows who influenced whom. This might be characterized as the number of nodes that gradually accept the information. Thinking of a node as either activated (i.e., having received the information and attempting to distribute it) or not is the easiest approach to explain the spreading process. Consequently, the propagation process may be understood as the activation sequence (specified in Definition 6) of a series of node activations distributed throughout the network.

Models created in the context of online social networks typically make the assumption that an individual is solely impacted by the behaviors of their connections. To clarify, They believe that information propagates via informational cascades and regard an OSN as a closed environment. This explains why the spreading cascade—described in Definition 7—is frequently used to describe the path a piece of information takes in the network, or the diffusion graph. To extract activation sequences from data, gather messages pertaining to the issue under study and arrange them in chronological order. Figure 5 provides an illustration of this concept. It tells you when and where a piece of information spread, but not how or why. So, in summary, Models that are able to depict and forecast the secret mechanism underlying diffusion are required. In this context, models can be divided into two categories: (i) explanatory models and (ii) predictive models. We describe these two groups and examine some exemplary work in each of them in the sections that follow.

4.1 Models of Explanation

Using a whole activation sequence as a starting point, explanatory models seek to deduce the underlying spreading cascade. With the use of these models, one may follow the trajectory that some data took.

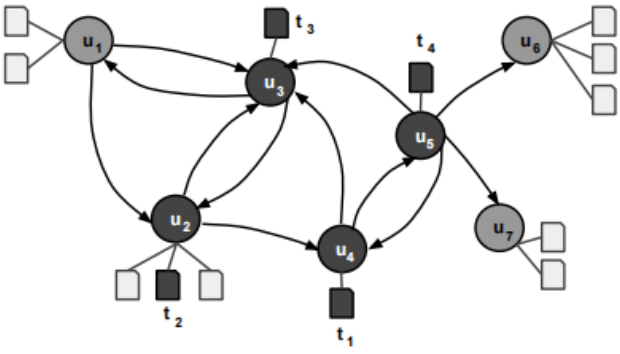

Figure 5: An OSN in which darker nodes took part in the diffusion process of a particular information.

The activation sequence can be extracted using the time at which the messages were published: [u4; u2; u3; u5], with t1 < t2 < t3 < t4. and are highly helpful in understanding the disseminationof information. In order to deduce the nature of the spreading cascade, Gomez et al. [15] suggest examining correlations in the timings at which nodes become infected. They also assume that active nodes have some degree of independent effect over each of their neighbors. Consequently, the likelihood that a single node has sent data to the disparity in their activation times is getting smaller. They create NETINF, an iterative method for determining the spreading cascade that maximizes the chance of observed data, based on sub modular function optimization.

A spatially discrete network of continuous, conditionally independent temporal events that occur at varying speeds is the model that Gomez et al. [14] propose to use to explain the diffusion process. A node's probability of infecting another at a specific dependent on infection periods and the rate of transmission between the two nodes, time is characterized by a probability density function. A convex maximum likelihood issue is

formulated and solved by the suggested algorithm, NETRATE, to estimate pairwise transmission rates and the diffusion graph [9]. The underlying network is assumed to remain static throughout these procedures. Given how quickly edges are added and removed, this is not a satisfying assumption on the topology of OSNs. That's why Gomez et al. [16] expand on NETRATE and suggest and present INFOPATH, a time-varying inference technique that estimates the structure and temporal dynamics of a network that is changing over time online using stochastic gradients.A data collecting bottleneck also exists as a result of technical and crawling API constraints.

Table 2: Summary of explanatory models w.r.t the nature of the underlying network, inferred properties and the ability of the method to work with incomplete data

reference	network		inferred properties			supports missing data
	static	dynamic	pairwaise transmission probability	pairwaise transmission rate	cascade properties	
NETINF	x		x		x	
NETRATE	x		x	x	x	
INFOPATH	x	x	x	x	x	
k-tree model	x				x	x

One way to get around this problem is to efficiently crawl data Choudhury and associates [7] examined the effects of data sampling approach on the finding of information diffusion in social media. They came to the conclusion—based on experiments conducted on Twitter data—that sampling techniques that take into account users' attributes like activity and localization as well as network topology can capture information diffusion with less error than naive approaches like random or activity-only based sampling. Creating specialized models that make the assumption that some data are absent is another strategy. Based on a ktree model, Sadikov et al. [41] devise a technique to estimate the parameters of the full spreading sequence given only a portion of the activation sequence. cascade's dimensions, for example.In Table 2, we provide a summary of the explanatory models surveyed. We describe the second class of models—predictive models—in the sections that follow.

4.2 Models of Predictability

Through the use of learning, these models seek to forecast the course of a certain diffusion process from temporal and/or spatial points of view within a given network.through historical dissemination traces. Graph and non-graph based techniques are the two development axes into which we divide the current models.

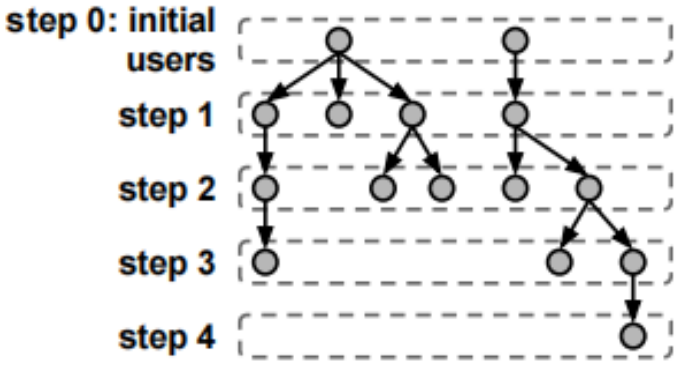

Figure 6: A spreading process modeled by Independent Cascades in four steps.

4.2.1 Methods based on graphs

In this category, Independent Cascades (IC) [13] and Linear Limit (LT) [17]. They concentrate on the structure of the process and presume that the diffusion is supported by a static graph structure. They are predicated on a directed network in which every node has the ability to activate or deactivate under the monotonicity assumption, meaning that activated nodes cannot deactivate. While each edge in the IC model needs to have a diffusion probability assigned to it, each node in the LT model needs to have an influence threshold and each edge needs to have its impact degree established. In all cases, the diffusion process begins at a group of initially activated nodes, referred to as early diffusion nodes, and moves iteratively in a synchronous manner along a discrete time-axis. Early Adoptors, Definition 8. a group of users who are the first to embrace a piece of knowledge and then start its spread.In the instance of IC, each time around, the freshly With the probability specified on the edge connecting them, active nodes attempt once to activate their neighbors. If the total of the influence degrees in the case of LT surpasses the individual influence threshold, the inactive nodes are activated by their activated neighbors at each iteration. Activations that are successful carry over to the following cycle. When an adjacent node cannot be connected or no further transmission is possible, the process terminates in both scenarios. Since LT is receiver-centric and IC is sender-centric, these two mechanisms represent two opposing viewpoints. A case of dissemination.Figure 6 displays a spreadin process modelled using IC. Models derived from such approaches and tailored for OSNs are described in detail below.With the process's commencement already observed, Galuba et al. [11] suggest using the LT model to forecast the diffusion graph. Their approach depends on factors including the likelihood that a user will adopt any knowledge, the degree of influence that pairs of users have over one another, and the virality of the information. The LT configurationis fitted using the gradient ascent method to optimize the parameters, describing the start of the diffusion process. LT is unable to replicate actual temporal dynamics, though. the proposal of asynchronous extensions, Saito et al. [42] loosen the synchronization assumption of conventional IC and LT graph-based models. The asynchronous independent cascades and asynchronous linear threshold, or AsIC and AsLT, respectively, operate iteratively along a continuous time axis and require the same parameters as their synchronous equivalents in addition to a time-delay parameter on each edge of the graph. The authors present a technique to determine the functional dependency of the model parameters from the attributes of the nodes, and the model parameters are defined in a parametric manner. They define the challenge as an updating algorithm and a maximum likelihood estimation problem that ensures the derivation of convergence. Nevertheless, they don't offer a workable solution and have only tested with artificial data. The propagation mechanism is also modeled by Guille et al. [19] as asynchronous separate cascades. They create the Time-Based Asynchronous Independent Cascades, or T-BaSIC, model, whose parameters are functions that depend on time rather than set numerical values. Using logistic regression, the model parameters are estimated using the properties of the social, semantic, and temporal nodes.

4.2.2 Methods that are not dependent on graphs

Non-graph-based methods have been around for a while and don't rely on the existence of a particular graph structure. primarily designed to simulate epidemiological processes. They divide nodes into multiple groups, or states, and concentrate on how the percentages of nodes within each class have changed over time. The two foundational theories are SIR and SIS [21, 34], where S stands for "susceptible," I for "infected" (i.e., assimilated the information), and R for recovered (i.e., refractory). Nodes in the S class migrate to the I class with a fixed probability β in both scenarios. Next, nodes in the I class in SIS switch to the S class with a fixed probability γ, but in SIR, they permanently switch to the R class. There are basic differential equations that express the fraction of nodes in each class. Both models presuppose that each node has an equal chance of being connected to another, meaning that connections within the population are formed at random. A straightforward and understandable SIS model with just one parameter, β, is proposed by Leskovec et al. [28]. All nodes are assumed to have an equal likelihood, β, of adopting the information, and nodes that

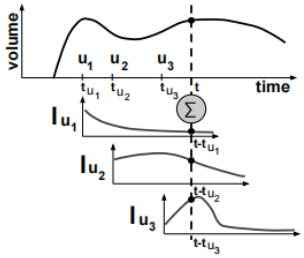

Figure 7: LIM forecasts the rate of diffusion by summing the influence functions of a given set of early adopters. Here, the early adopters are u1, u2 and u3 whose respective influence functions are Iu1, Iu2 and

Iu3.accept the information and become vulnerable at the following time step (i.e., $\gamma = 1$). In actual social networks, influence is not dispersed equally among all nodes, hence this is a strong assumption.

Consequently, more intricate modeling that considers this feature must be developed.
Starting from the premise that individual nodes' influence controls the spread of information, Yang et al. [50] work. Predicting information's temporal dynamics is the main goal of the approach.diffusion is the rate at which a piece of information spreads, or the number of nodes that adopt the information over time, and it takes the shape of a time series. In a Linear Influence Model (LIM) they develop, the total rate of diffusion is determined by the influence functions of individual nodes. Through the use of the Reflective Newton Method, a non-negative least squares problem may be solved to estimate the influence functions in a non-parametric manner [8]. A set of early adopters' rate of diffusion and their activation time are forecasted by LIM, as shown in Figure 7.
A model based on partial differential equations (PDEs) is proposed by Wang et al. [48] to forecast the diffusion of information injected into the network by a specific node. To be more specific, topological and dynamics with time. Here, the network topology is solely taken into account in terms of the separation between each node and the source node. A logistic equation that represents the density of affected users at a specific time and distance from the source provides the process dynamics. By incorporating some spatial knowledge, the problem can be formulated as easily using that specification of the network topology as it can be for traditional non-graph-based methods.

Table 3: Summary of diffusion prediction methods, distinguishing graph and nongraph based approaches w.r.t incorporated dimensions and mathematical modeling.

reference	dimension(s)			basis		mathematical modeling	
	social	time	content	graph based	non-graph based	parametric	non-parametric
LT-based	x		x	x		x	
AsIC, AsLT	n/a	n/a	n/a	x		x	
T-BaSIC	x	x	x	x		x	
SIS-based		x			x	x	
LIM	x	x			x		x
PDE	x	x			x	x	

The Cubic Spline Interpolation method [12] is used to estimate the model's parameters.The surveyed predictive models are compiled here.
Table 3. The function of nodes in the propagation process and techniques for identifyingsignificant spreaders are covered in the section that follows.

5. Identifying information spreaders that are influential
For information to travel effectively throughout a network, it is essential to identify its most powerful spreaders. Targeting well-known people, for example, can maximize the results of a social media campaign. who is able to start significant adoption cascades. The several approaches of determining the relative significance and effect of every node in an online social network are briefly discussed in this section.

Definition 9
(K-Core). Let G be a graph If f H is a sub-graph of G, then $\sigma(H)$ will represent H's lowest degree. Every node in H is therefore adjacent to at least $\sigma(H)$ other nodes in H. A k-core of G is defined as H being a maximal connected (induced) sub-graph of G with $\sigma(H) >= k$ [45].It may be shown by Kitsak et al. [25] that the most effective spreaders are not always the most connected individuals in the interconnect. As revealed by the k-core decomposition study, they discover that the spreaders with the highest efficiency are those situated in the network's core [45]. as outlined in Definition 9. The basic idea of the k-core decomposition is to give each node

a core index, ks, so that the nodes with the highest values are found in the network's center and the nodes with the lowest values are found on its edges. As a result, the network's core is made up of the deepest links. The k-shell decomposition results on the Twitter network are noticeably skewed, as noted by Brown et al. [5]. Therefore, in order to generate fewer and more significant k-shell values, they suggest modifying the algorithm to employ a logarithmic mapping.
The well-known PageRank algorithm [35] is suggested to be used by Cataldi et al. [6] to evaluate the influence's global distribution. An individual node's PageRank value is correlated with the likelihood of the node being visited in a random walk of the social network, where the nodes make up the collection of states for the random walk. These techniques just take use of the network's structure; they don't consider other crucial characteristics like the characteristics of the nodes and how they process data. Romero et al. [38] build a graph-based method which, like the well-known HITS algorithm, IP (i.e. Influence-Passivity), based on the observation that the majority of OSNs members are passive consumers of information. the ratio at which users forward information determines the passivity score for each user. Nonetheless, it is impossible for a single person to have a universal effect, and important network members typically only have power in one or a few particular fields of expertise. Pal et al. [36] so devise a topic-sensitive, non-graph-based approach. For the purpose of characterizing the network members, they define a set of nodal and thematic features. To determine which individuals are the most important and knowledgeable about a certain issue, they rank nodes using a within-cluster ranking process over this feature space using probabilistic clustering.TwitterRank, a topic-sensitive variant of the Page Rank algorithm specifically designed for Twitter, is also developed by Weng et al. [49].

Table 4: Summary of influential spreaders identification methods distinguishing graph and non-graph based approaches w.r.t incorporated dimensions.

reference	graph based	incorporated dimension(s)	
		users' features	topic
k-shell decomposition	x		
log k-shell decomposition	x		
PageRank	x		
Topic-sensitive PageRank	x		x
IP	x	x	
Topical Authorities		x	x
k-node set	x		

as the set of nodes that were first active, then send that set to IC or LT. Using a greedy hillclimbing technique based on submodular functions, they offer an approximation for thisoptimizationproblem.
Table 4 summarizes the strategies for assessing influence that were surveyed.

Conversation
We reviewed typical and cutting-edge techniques for information diffusion analysis in online social networks in this study, covering a variety of ranging from strategies for detecting hot topics to diffusion modeling approaches, which include ways to identify important spreaders. The taxonomy of the many strategies used to deal with these problems is shown in Figure 8. We now address their deficiencies and associated unresolved issues.

6.1 Identifying Trending Subjects
The identification of bursts is a prerequisite for the recognition of popular subjects from the stream of messages sent by OSN members. Primarily, there are two methods to look for these patterns: (i) term frequency analysis, or (ii) social interaction frequency analysis.
This region, the following challenges certainly need to be addressed:
Terminology and scalability of topics. It's clear that different approaches define the same subject differently. In Peaky Topics, for example, a topic is simply converted to a word. Its benefit is that it's a cheap complicated problem, however the outcome that is generated is hardly interesting. However, because of its tremendous

complexity, OLDA cannot be applied on a broad scale. Instead, it defines a subject as a distribution over a set of words. As a result, a necessity searching for fresh techniques that could maintain efficiency while yielding understandable outcomes. We point out two potential approaches to do this: (i) developing new, scalable algorithms, or (ii) enhancing existing algorithms through the use of distributed systems (like Hadoop). aspect of society. Additionally, popular subject detection, like TSTE, which depends on the Page Rank method, could be enhanced by utilizing burstiness and people authority. Nevertheless, that potential hasn't been thoroughly investigated yet

Data intricacy: At the moment, the emphasis is on the textual material that is shared on social networks.

But increasingly commonly, users trade other kinds of data, like pictures, movies, URLs leading to those things on the Internet, etc. In order to create a comprehensive topic identification system, this issue needs to be thoroughly taken into account and integrated into the overall effort.

6.2 Representing Diffusion of Knowledge

Explanatory and predictive models are the two categories into which we separate them. With regards to predictive models, on One approach, which is not based on graphs, is constrained in that it only predicts the evolution of the rate at which information diffuses worldwide, ignoring the network's topology. Graph-based techniques, on the other hand, can forecast who will impact who. But when the network is implicit or unknown, they cannot be utilized. Studying information dispersion requires taking into account more practical limits, despite the fact that a lot of work has been done in this field. It is especially necessary to address the following problems:

Definition 10 (Closed World)

The closed world assumption holds that information can only propagate from node to node via the network edges and that nodes cannot be influenced by external sources.

limited world perspective. The most important thing to note when modeling the spread of knowledge is that every approach discussed operates under the premise of a closed world, as stated in Definition 10.To put it another way, they believe that individuals can onlybe affected by other network users, and informational cascades cause that information to spread. But the majority of those seen distributing pro-

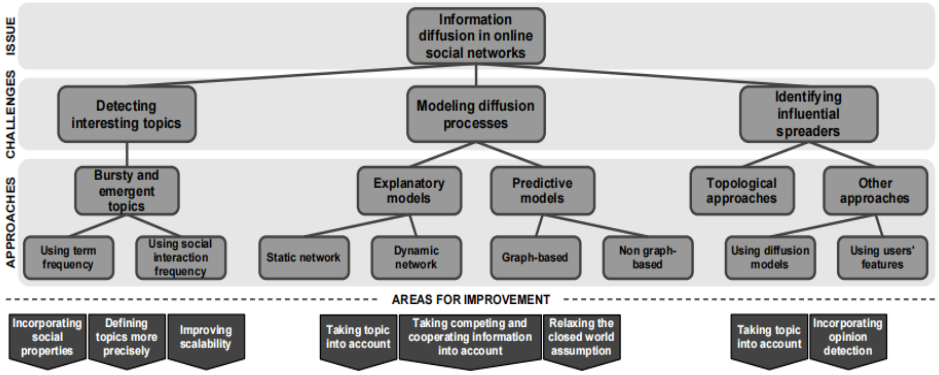

Figure 8: The above taxonomy presents the three main research challenges arising from information diffusion in online social networks and the related types of approaches, annotated with areas for improvement.

Social influence is not the exclusive source of cesses in OSNs. Recent study on Twitter by Myers et al. [32] shows that information tends to move across the network, disproving the closed-world assumption. The analysis demonstrates that only 71% of the information volume on Twitter can be attributable to internal influences, with the remaining 29% coming from outside events and influences. As a result, they offer a model that uses hazard functions to quantify the degree of external exposure and influence [10]. Aligning users' profiles on several social networking sites might be one method to dispel this belief. This would allow for the simultaneous observation of information dissemination across several platforms (subject to the availability of data). Proposals to de -anonymize social networks have been made in some studies to solve issues of this nature [33].

Diffusion processes that compete and work together

Furthermore, the research stated above are predicated on the idea that diffusion processes are autonomous, that is, that eachpiece of information disseminates independently. According to Myers et al. [31], propagation processes Work with and against each other. While cooperating contagions aid in each other's adoption, competing ones reduce each other's likelihood of spreading.They put out a model that measures the interplay between various spreading cascades. It forecasts diffusion probabilities that, on average, differ by 71% from

what a diffusion process that is totally independent would have. We think that this knowledge needs to be taken into account by models.

Subject-specific modeling
It is imperative that prediction models exhibit topic-sensitivity. Romero et al.'s [39] analysis of Twitter revealed notable variations in the mechanisms behind the spread of information among subjects. Specifically, they have noticed that politically contentious information is especially persistent, with repeated exposures maintaining abnormally large marginal effects on adoption. This supports the complex contagion principle, which states that repeated exposures to an idea are especially important when the idea is contentious or controversial.

Network Dynamics
OSNs are extremely dynamic structures, which is a final point worth making.However, the presumption that the network stays static over time underlies the majority of the work now in existence. Enhancing the accuracy of predictions may begin with integrating link prediction. [20] provides a more thorough analysis of the literature on this subject.

6.3 Recognizing Key Disseminators
This problem can be solved in a number of ways, from pure topological techniques like kshell decomposition or HITS to textual clustering-based techniques like hybrid approaches an IP that blends the properties of nodes with the HITS algorithm. Since there isn't a one universal influencer—as was previously mentioned—topic-sensitive approaches have also been created.

Opinion identification
The concepts of opinion and influence are closely related. Recent years have seen the emergence of numerous studies on this topic, all attempting to automatically discern opinions or sentiment from the data corpus. It is our opinion that In the perspective of information dissemination, it could be interesting to add this kind of activity. It appears that there is curiosity in the work that has evolved on the diffusion of opinions themselves [29] to combine these strategies.

6.4 Applications
We might note that implementations are rarely made available for reuse in the field of online social network dynamics studies. Furthermore, it is challenging to assess or compare current methods since differing programming languages and requirements for the input data formatting affect the available implementations. SONDY [18] plans to make it easier to deploy and disseminate data mining approaches for online social networks. It's an open-source tool that performs data preprocessing functions and applies some of the techniques discussed in this work to identify prominent spreaders and identify topics. It offers visuals for network structure and topic trends, and it has an intuitive user interface.

REFERENCES
[1] L. AlSumait, D. Barbar´a, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In ICDM ’08, pages 3–12, 2008.
[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In KDD ’08, pages 7–15, 2008.
[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In WWW ’12, pages 519–528, 2012.
[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.
[5] P. Brown and J. Feng. Measuring user influence on Twitter using modified k-shell decomposition. In ICWSM ’11 Workshops, pages 18–23, 2011.
[6] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In MDMKDD ’10, pages 4–13, 2010.
[7] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In ICWSM ’10, pages 34–41, 2010.
[8] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. SIAM J. on Optimization, 6(4):1040–1058, Apr. 1996.
[9] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, sep 2012.
[10] R. C. Elandt-Johnson and N. L. Johnson. Survival Models and Data Analysis. John Wiley and Sons, 1980/1999.

[11] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitters - predicting information cascades in microblogs. In WOSN ’10, pages 3–11, 2010.
[12] C. F. Gerald and P. O. Wheatley. Applied numerical analysis with MAPLE; 7th ed. Addison-Wesley, Reading, MA, 2004.
[13] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001.
[14] M. Gomez-Rodriguez, D. Balduzzi, and B. Sch¨olkopf. Uncovering the temporal dynamics of diffusion networks. In ICML ’11, pages 561–568, 2011
[15] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In KDD ’10, pages 1019–1028, 2010.
[16] M. Gomez-Rodriguez, J. Leskovec, and B. Sch¨okopf. Structure and dynamics of information pathways in online media. In WSDM ’13, pages 23–32, 2013.
[17] M. Granovetter. Threshold models of collective behavior. American journal of sociology, pages 1420–1443, 1978.
[18] A. Guille, C. Favre, H. Hacid, and D. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In SIGMOD ’13, (demonstration) 2013.
[19] A. Guille and H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In WWW ’12 Companion, pages 1145–1152, 2012.
[20] M. A. Hasan and M. J. Zaki. A survey of link prediction in social networks. In Social Network Data Analytics, pages 243–275. Springer, 2011.
[21] H. W. Hethcote. The mathematics of infectious diseases. SIAM REVIEW, 42(4):599–653, 2000. [22] P. N. Howard and A. Duffy. Opening closedregimes, what was the role of social media during the arab spring? Project on Information Technology and Political Islam, pages 1–30, 2011.
[22] A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3):248–260, 2009.
[23] D. Kempe. Maximizing the spread of influence through a social network. In KDD ’03, pages 137–146, 2003.
[24] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identification of influential spreaders in complex networks. Nature Physics, 6(11):888–893, Aug 2010. [26] J. Kleinberg. Bursty and hierarchical structure in streams. In KDD ’02, pages 91–101, 2002.
[25] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In KDD ’09, pages 497–506, 2009.
[26] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In SDM ’07, pages 551–556, (short paper) 2007.
[27] L. Li, A. Scaglione, A. Swami, and Q. Zhao. Phase transition in opinion diffusion in social networks. In ICASSP ’12, pages 3073–3076, 2012.
[28] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple semantics in topic detection and tracking. Inf. Retr., 7(3-4):347–368, Sept. 2004.
[29] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In ICDM ’12, pages 539–548, 2012
[30] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In KDD ’12, pages 33–41, 2012.
[31] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In SP ’09, pages 173–187, 2009.
[32] M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45:167–256, 2003.
[33] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In WWW ’98, pages 161–172, 1998.
[34] A. Pal and S. Counts. Identifying topical authorities in microblogs. In WSDM ’11, pages 45–54, 2011.
[35] E. M. Rogers. Diffusion of Innovations, 5th Edition. Free Press, 5th edition, aug 2003.
[36] D. Romero, W. Galuba, S. Asur, and B. Huberman. Influence and passivity in social media. In ECML/PKDD ’11, pages 18–33, 2011.
[37] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In WWW ’11, pages 695–704, 2011.
[38] L. Rong and Y. Qing. Trends analysis of news topics on Twitter. International Journal of Machine Learning and Computing, 2(3):327–332, 2012.
[39] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In WSDM ’11, pages 55–64, 2011.
[40] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda. Learning diffusion probability based on node attributes in social networks. In ISMIS ’11, pages 153–162, 2011.
[41] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage., 24(5):513–523, 1988.

[42] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1986.
[43] S. B. Seidman. Network structure and minimum degree. Social Networks, 5(3):269 – 287, 1983.
[44] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In CSCW ’11, pages 355–358, (short paper) 2011.
[45] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering emerging topics in social streams via link anomaly detection. In ICDM ’11, pages 1230–1235, 2011
[46] F. Wang, H. Wang, and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In ICDCS ’12 Workshops, pages 133–139, 2012.
[47] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitters. In WSDM ’10, pages 261–270, 2010.
[48] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In ICDM ’10, pages 599–608, 2010.